

5 CONCLUSIÓN Y TRABAJO A FUTURO

En esta tesis se presentó un extractor terminológico basado en conocimiento estadístico, junto con un sistema que emplea la enciclopedia Wikipedia como recurso léxico para validar los candidatos a término. Con ello se cumplió el objetivo principal y el objetivo secundario:

- Se desarrolló un sistema extractor terminológico basado en conocimiento estadístico, más específicamente, en Term Frequency – Inverse Frequency Document (TF-IDF). El cual trae la mayor cantidad de términos posibles para obtener mejores resultados de cobertura.
- Se empleó Wikipedia como un recurso léxico para llevar a cabo la validación de cada uno de los candidatos a término que se obtuvieron del extractor terminológico.
- Se obtuvieron tres listas de términos validadas para las materias de matemáticas y ecología de bachillerato, así como la de matemáticas de educación primaria.

La realización de este trabajo de tesis permitió llegar a varias conclusiones que a continuación se describirán siguiendo la estructura de la tesis, es decir, primeramente se comenzará con la del corpus COCIEM, luego la del extractor terminológico, posteriormente con la de la validación empleando Wikipedia y finalmente la de los resultados obtenidos.

El Corpus de textos científicos en español de México (COCIAM) ha sido un gran corpus de prueba para llevar a cabo la extracción terminológica y su validación empleando Wikipedia. La razón de lo anterior es que abarca una gran cantidad de materias científicas, que van de nivel primaria a nivel bachillerato, ricas en términos que pueden ser validados empleando recursos léxicos no tan especializados o de alto nivel. Sin embargo, considero que tiene sus deficiencias como palabras mal acentuadas, la inclusión de símbolos extraños o errores en la estructura del libro debido al escaneado (partes del texto a doble columna, discontinuidad por cuadros del estilo ¿sabías qué?, etcétera), los cuales pudieron ser prevenidos en el momento del escaneado.

Con respecto al extractor terminológico ciertamente existen deficiencias en él, debido a que no se pueden obtener ciertos términos, como lo son los que empiezan con artículos, por

ejemplo “El Niño”, el cual es un fenómeno climatológico, lo anterior debido a la limpieza que se llevó a cabo en las listas de candidatos a término antes del cálculo del IDF que consistió en la eliminación de n-gramas con palabras vacías en sus extremos. Asimismo, la clasificación de los candidatos a término por los pesos de TF-IDF no fue tan buena como se hubiera esperado, pues la mayoría de los candidatos quedaban en pesos medios y bajos, y no altos, posiblemente por el alto número de palabras externas a las materias del COCIEM por su carácter educativo, como lo son el empleo de ejemplos que usan la vida cotidiana, la narración de diálogos o historias con nombres de personas para dar a conocer un tema o la inclusión de cuadros del estilo de ¿sabías qué? que describen información adicional o relacionada con otras materias para mejorar la comprensión del tema. Sin embargo, pienso que el extractor cumplió su cometido trayendo la mayor cantidad de candidatos a término posibles para tratar de tener la mayor cobertura; además creo que el uso de los pesos de TF-IDF puede ser de gran ayuda si se lleva a cabo un análisis más profundo de los pesos que otorga para calcular umbrales mucho más específicos, y por tanto separar de mejor manera los posibles candidatos a términos de los que no tienen posibilidad de serlo, ya sea por su poca relevancia en el documento o por su poca repetición en el corpus de análisis.

La validación de los candidatos a término fue un método que no sólo permitió validar los resultados obtenidos del extractor terminológico, sino que complementó en una gran medida el uso de un método estadístico para la extracción de términos, debido a que descartó un gran número de candidatos a término que ciertamente no eran términos por la estructura morfosintáctica que tenían, algo que sólo se puede realizar empleando conocimiento lingüístico en la etapa de extracción.

Los resultados, en el caso de ecología de bachillerato y matemáticas de primaria son buenos, en el caso de matemáticas de bachillerato no son tan buenos como los anteriores, pero por la prueba realizada con otra lista de evaluación se muestra que sus resultados no tan buenos fueron principalmente por la lista de evaluación original más que por la validación realizada por Wikipedia.

Ciertamente, los resultados que se obtuvieron pudieron ser mejores, pero algunos de los errores que ocurrieron en la validación empleando Wikipedia son intrínsecos a la estructura de la enciclopedia y no tanto al sistema extractor de términos. En este caso existen

ejemplos en donde la estructura enciclopédica de Wikipedia afecta los resultados obtenidos, por tener categorías como “Cultura por país”, “Biografías por actividad”, entre otras, que unen ciertos temas con categorías que no están totalmente relacionadas, como los mostrados en la sección 4.4.1. De igual forma, al ser Wikipedia una enciclopedia creada por las personas, existen errores u omisiones en la asignación de las categorías, los cuales afectan el proceso para el cálculo del coeficiente de dominio. Es necesario, de igual forma, incluir en esta parte lo de la desambiguación de términos, pues aunque en algunos casos obtenía el término correcto, en algunos otros no lo hacía, debido al método elegido (término con máximos coeficientes de dominio) y no tanto a Wikipedia en sí. A pesar de todo lo anterior, considero que Wikipedia es un excelente recurso que puede ser empleado en la validación de candidatos a término porque abarca una gran cantidad de materias (o categorías) y de temas que no se encuentran frecuentemente en un solo recurso; además de que Wikipedia está en constante crecimiento y actualización. Solamente, es necesario que se estudie más a fondo la estructura de Wikipedia para omitir ciertas características propias de ella que a veces afectan los resultados y mejorar la manera en que se desambiguan los términos.

Existe una gran cantidad de puntos a trabajar a futuro para mejorar tanto la extracción terminológica como la validación empleando Wikipedia. La primera tarea sería la de mejorar el extractor terminológico, ya sea empleando otra métrica o mezclando varias de ellas para tener distintas clasificaciones y obtener mejores candidatos a término. La segunda tarea sería desarrollar nuevas fórmulas para el cálculo del coeficiente de dominio, para ver si se pueden obtener mejores resultados. Asimismo, como tercera tarea sería encontrar una mejor manera de encontrar los términos correctos que son ambiguos, pues el método que se empleó en esta tesis no es del todo confiable. La cuarta tarea a realizar a futuro sería llevar a cabo un análisis más profundo de la estructura de Wikipedia para encontrar, y omitir posteriormente en el análisis, las categorías que pueden causar casos erróneos en la validación de los candidatos a término. La quinta tarea sería emplear la misma Wikipedia como un recurso que convierta los resultados lematizados a su forma correcta de mostrarse, es decir, que haya concordancia en género y número y no se muestren casos como “fracción impropio” sino como “fracción impropia”. Finalmente, la sexta tarea sería aplicar el método de extracción de términos y su validación empleando Wikipedia a todas las materias que pertenecen al COCIEM.